A SIMULATION STUDY TO COMPARE THE POWER OF NORMALITY TESTS

Péter Erdélyi, Róbert Rajkó

Institute of Process Engineering, University of Szeged, H-6725 Szeged, Moszkvai krt. 5-7., Hungary
e-mail: erd_pet@fastmail.jp

Abstract
A common assumption of many statistical procedures during data analysis is that the data is normally distributed. Several statistical tests have been developed for the determination of the validity of this assumption. Now, the question is: which is the most powerful? We performed an extensive Monte Carlo simulation study to answer this question. We found that compared to six other tests, the Shapiro–Wilk test performs the best under most conditions.

Introduction
One of the most frequently made assumptions during data analysis (for example when performing a t-test, an F-test, an analysis of variance, or tests for the regression coefficients in a regression analysis) is that the data is normally distributed. [1]
There are graphical methods to determine the normality of the data (Q-Q plots, etc., which we will not discuss here since these are less reliable, subjective), and there are also dozens of more objective, formal statistical tests. The mere fact, that there are so many alternatives suggests that this is a hard problem without an exact, perfect solution, and also, it raises the obvious question: which one should we use, which is the best test?
The tests we have studied are based on frequentist inference: they are hypothesis tests, and the data set is tested against the null hypothesis that it is normally distributed. Therefore, it can commit either of the well-known two types of error: it incorrectly rejects the null hypothesis (type I error, “false positive”), or incorrectly fails to reject it (type II error, “false negative”).
If we generate a large number of normally distributed samples, carry out a normality test on each, then count the cases when the test incorrectly rejects the null at a given significance level (α) we can calculate the proportion of type I errors. This value is called the *size* of the test. The significance level chosen by the user is the upper bound of the size, if the test works correctly.
The *power* of a test ($1-\beta$) is the complement of the type II error rate (β), which can be determined similarly: by generating samples of non-normal distributions, and count the cases of the test failing to reject the null hypothesis of normality.

Experimental
We performed an extensive Monte Carlo simulation study in the R statistical programming environment.[2] The tests compared were the Shapiro–Wilk test (SW for short in the following), provided by the R “stats” package, which is part of the base distribution, so it can be considered as the default normality test in R, the Jarque–Bera test (JB) from the “moments” package, and five more tests from the “nortest” package: the Shapiro–Francia (SF), Anderson–Darling (AD), Cramér–von Mises (CM) Lilliefors (LI), and Pearson's X^2 (PE) tests.
We generated samples from the standard normal distribution and 17 alternative (non-normal) distributions which we divided to two groups: the symmetricals (i.e. with a skewness of zero,

like the normal distribution), and the asymmetricals (with nonzero skewness) to see if any tests perform better against one type of alternative distribution than the other.

Symmetric distributions	**Asymmetric distributions**
Laplace (μ=0, σ=0.5)	X^2 (k=1), (k=20)
Logistic (μ=0, s=0.5)	Exponential (λ=0.5), (λ=1)
Student's t (ν=1), (ν=20)	Poisson (λ=10)
Beta (α=0.5, β=0.5), (α=2, β=2)	Log-normal (μ=0, σ=1), (μ=0, σ=0.1)
Uniform (a=0, b=1)	Weibull (λ=0.5, k=1), (λ=10, k=1)
Binomial (n=20, p=0.5)	

The sample sizes were: 10, 15, 20, 25, 30, 35, 50, 100, 200, 500, the number of samples were 1.000.000 each of the ten sizes. We then evaluated all seven tests on every sample at three significance levels: 0.01, 0.05, 0.1, and calculated their power and size as explained in the introduction.

Results and discussion

Considering the large amount of generated data, we summarize the results on Figures 1–3. The best performing test is uniformly the Shapiro–Wilk test, closely followed by the Anderson–Darling, the Shapiro–Francia and the Cramér–von Mises test.

As seen on Figure 1, The Shapiro–Wilk test holds the set significance level almost perfectly. The Shapiro–Francia test exceeds it slightly, but that improves with the increasing sample size. The other tests' type I error rate stays under the required level, which is not optimal, but not erroneous. The behavior of the Pearson test is almost pathological, for reason unknown.

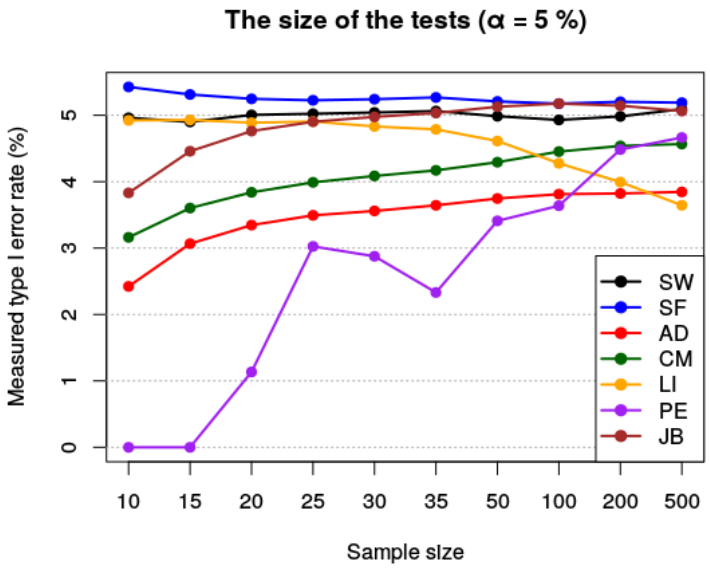

Figure 1. The size of normality tests as a function of sample size at a significance level of 5%. (The horizontal axis is not to scale.)

As expected, the power of every test is increasing with the sample size – there's more information to base the decision on (Figure 2). The standard deviation is generally high, because we are averaging the results for several alternative distributions, some of them are easy to detect (e.g. Student's t with 1 degree of freedom), and also some harder cases (e.g. Student's t with 20 degrees of freedom). The interesting, upside down "U" shape of the SD vs.

sample size curves is due to the fact that the power at low sample sizes is usually close to its lower bound (0%), at high sample sizes it's close to its upper bound (100%), and there's simply less room for the value to vary.

If we look at the symmetrical and asymmetrical distributions separately (Figure 3), the trend is basically the same: the SW, AD, SF and CM tests have the highest power. In the asymmetrical cases the average power is higher, and also the difference between the tests is smaller, because the skewed distributions are easier to detect, they are more different from the normal distribution.

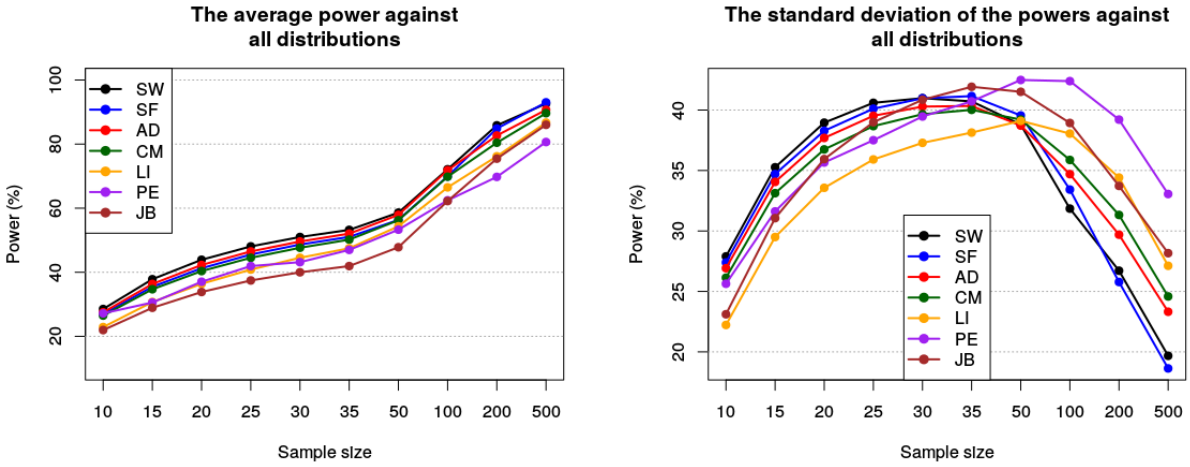

Figure 2. The average and the standard deviation of the power of normality tests against all studied distributions in the function of sample size. (The horizontal axes are not to scale.)

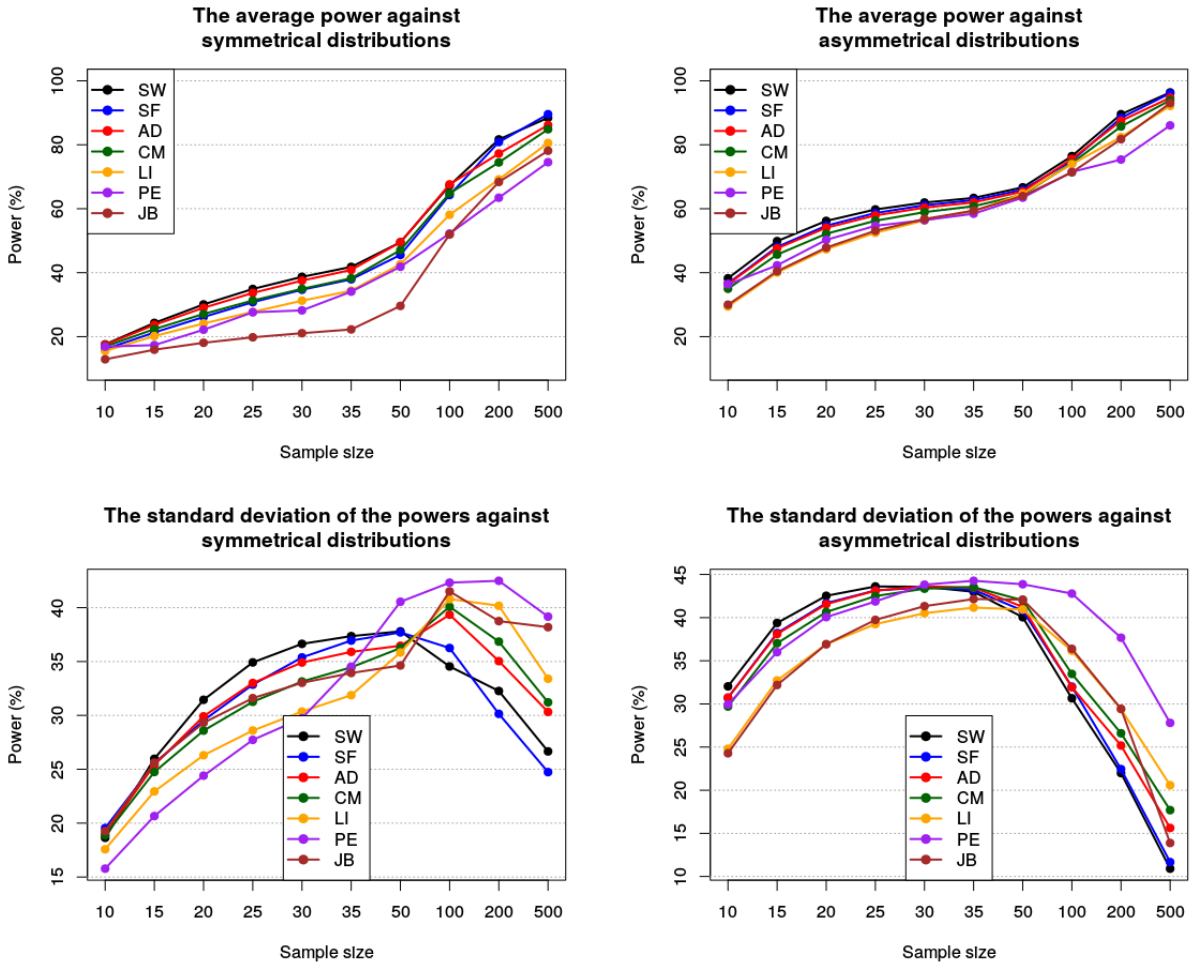

Figure 3. Comparison of the average and the standard deviation of the power of normality tests against symmetric and asymmetric distributions. (The horizontal axes are not to scale.)

Conclusion

We compared the power of seven normality tests available in R statistical computing environment, and found that the Shapiro–Wilk test performs the best against a large variety of alternative distributions both at small and large sample sizes. This test is part of the base R distribution as a de facto default, and it seems the developers have chosen wisely, there is no need to install any other packages.

References

[1] Thode, Henry C.: Testing for normality. Vol. 164. CRC press, 2002. p. 1.

[2] R Core Team, R: "A Language and Environment for Statistical Computing." R, Foundation for Statistical Computing, Vienna, Austria, 2015. https://www.R-project.org